ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Robust object segmentation by adaptive metrics in Generalized LVQ

Alexander Denecke[1,2], Heiko Wersing[2], Jochen J. Steil[1], Edgar Körner[2]

1- Bielefeld University - Neuroinformatics Group
P.O.-Box 10 01 31, D-33501 Bielefeld - Germany
adenecke@techfak.uni-bielefeld.de

2- Honda Research Institute Europe
Carl-Legien-Str. 30, 63073 Offenbach/Main - Germany

Abstract.
We investigate the effect of several adaptive metrics in the context of figure-ground segregation, using Generalized LVQ to train a classifier for image regions. Extending the Euclidean metrics towards local matrices of relevance-factors does not only lead to a higher classification accuracy and increased robustness on heterogeneous/noisy data, but also figure-ground segregation using this adaptive metrics enables a considerably higher recognition performance on segmented objects of real image data.

1 Introduction

Segregating a currently attended object from the surrounding background is fundamental for research on object learning, recognition, and interaction under general environment conditions (Fig. 1). If one cannot rely on foreground detection, an initial hypothesis about the object can be derived, for example, from depth estimation [1], information available from saliency-maps [2], or top-down information about object parts [3]. The problem is to extract the relevant object regions from such imprecise hypotheses for further processing. The Adaptive Scene-Dependent Filter (ASDF) [4] addresses the problem by over-segmenting the image into homogeneous regions and selecting the segments which match the hypothesis. Extending the ASDF by stating figure-ground segregation as a binary classification problem, we use Generalized Learning Vector Quantization (GLVQ [5]) to adapt a prototype-based classifier for figure and ground. When using a prototype-based representation, clustering and classifying image regions on the basis of similarity crucially depends on the underlying metrics. For GLVQ several extensions of the Euclidean metrics are available [6, 7] which offer additional feature and prototype-specific weighting factors, taking into account feature discriminability and correlations between them. Those so-called relevance-factors are online adapted with gradient descent together with the LVQ-network weights. By comparing the adaptive metrics, we show that the introduction of prototype-specific matrices of relevance-factors is capable of achieving a large gain in segmentation quality enhancing object learning and recognition. Compared to the ASDF, this method offers the advantage to automatically determine the best discriminating feature dimensions for object segmentation, and additionally relaxes a priori assumptions on object position and segment selection.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

2 Method

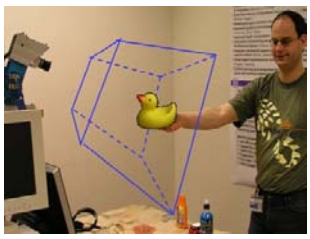


Figure 1: General setup showing the peripersonal space for object learning [8].

Our current scenario for object learning consists of a person presenting objects to a stereo-camera system. The pan-tilt stereo-camera head is controlled by a three-layered attention system to enable unconstrained learning. To localize, track and centre the object in view (translation invariance), the concept of peripersonal space is used [8], which defines the behaviourally relevant parts of the scene (bounding box in Fig. 1). From the blob-detection in the peripersonal space a square region of interest is defined (ROI) and normalized to a size of 144×144 pixels (size invariance). The depth information within this ROI is used as an initial object hypothesis $\mathcal{H}$.

Problem: Extracting 3D information from 2D images in general is an ill-posed problem, resulting in solutions with coarse approximations of the object by the depth estimation. Therefore, extracting the relevant object parts from this hypothesis is complicated by partially overlapping feature-clusters due to the noisy character of the hypothesis itself, as well as similar colors in regions of figure and ground. Formally the input data consist of M feature maps $\mathcal{F} := \{F_i | i = 1..M\}$ and here we use 6 feature-maps selected from RGB, HSV image and pixel position ($F_1^{x,y} = R^{x,y}$, $F_2^{x,y} = G^{x,y}$, $F_3^{x,y} = B^{x,y}$, $F_4^{x,y} = V^{x,y}$, $F_5^{x,y} = x$, $F_6^{x,y} = y$). The stack of maps is represented by a set of vectors $\vec{\xi}^{x,y} \in \mathbb{R}^M, 1 \leq x, y \leq 144$. We assume an unknown ground truth map $\mathcal{G}$, which defines for every pixel (x, y) the membership of $\vec{\xi}^{x,y}$ to figure $\mathcal{G}^{x,y} = 1$ or ground $\mathcal{G}^{x,y} = 0$ with respect to the attended object. The goal is to approximate $\mathcal{G}$ by another binary map $\mathcal{A}$ using the initial hypothesis $\mathcal{H}$ (also a binary map) and the similarity-information provided from the feature-maps $\mathcal{F}$. Finally the overlap with the *real* segmentation (see Fig. 2) must be increased $G(\mathcal{A}) > G(\mathcal{H})$, measured by the similarity function $G(\cdot)$ which is the ratio of intersection and union of $\mathcal{G}$ and $\mathcal{A}$

$$G(\mathcal{A}) := 1 - \frac{\sum_{x,y} |\mathcal{A}^{x,y} - \mathcal{G}^{x,y}|}{\sum_{x,y} \mathcal{A}^{x,y} + \sum_{x,y} \mathcal{G}^{x,y}}, \; \mathcal{H}^{x,y}, \mathcal{A}^{x,y}, \mathcal{G}^{x,y} = \begin{cases} 1 & for\ foreground \\ 0 & for\ background \end{cases}$$

Adaptive Scene-Dependent Filter: For the ASDF, $\mathcal{H}$ are the non skin-coloured areas (filtered in a separate processing stream for skin color detection) from a superposition of the depth-map with a position and size prior (circular map [4]). To build $\mathcal{A}$ and to extract the relevant object parts from $\mathcal{F}$ using $\mathcal{H}$, the current ASDF implementation basically consists of two steps (see [4] for details). After pre-processing the feature maps $F_i^{x,y} \leftarrow f_i \cdot (F_i^{x,y}/\sigma_i^2)$ by scaling with their variance $1/\sigma_i^2$ and a feature-specific a priori weighting factor f_i, a vector quantization is performed to segment the image. A modified *K-means* clustering is used to approximate the clusters in the data (homogeneous regions in the image) by a fixed set of N prototypes $\mathcal{P} := \left\{\vec{w}^p \in \mathbb{R}^M | p = 1..N\right\}$.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

After adapting the prototypes with Hebbian learning while randomly choosing samples $\vec{\xi}^{x,y}$, the image is partitioned into N segments (binary maps) V_p by assigning all feature vectors $\vec{\xi}^{x,y}$ (i.e. pixels) independently, to the prototype whose weight-vector has the smallest Euclidean distance $d(\vec{\xi},\vec{w}) = \|\vec{\xi}-\vec{w}\|$ to $\vec{\xi}$.

$$V_p^{x,y} := \begin{cases} 1 & if\ d(\vec{\xi}^{x,y},\vec{w}^p) < d(\vec{\xi}^{x,y},\vec{w}^r), \forall r \neq p, \{r,p\} \in \mathcal{P}, \\ 0 & else. \end{cases}$$

Finally, $\mathcal{A} = \sum_p^N V_p$ is constructed heuristically by using a subset of activation maps V_p, each of which shows a sufficient overlap with the initial hypothesis $\mathcal{H}$. Additionally, a temporal integration by reusing the old prototypes to initialize the network on the following image increases stability and compensates a reduced number of adaptation steps.

2.1 Generalized Learning Vector Quantization with Relevance-factors

Alternatively to unsupervised clustering one can state the task of object extraction as a binary classification problem and use learning (i.e. supervised) vector quantization to adapt class-specific prototypes. In both cases similarity-based clustering and classification crucially depends on the underlying metrics. Extending the Euclidean metrics used by ASDF and Generalized LVQ (GLVQ [5]) by introducing a relevance-factor for each feature dimension (Generalized *Relevance* LVQ (GRLVQ) [6]) leads to the squared weighted Euclidean metrics $\|\vec{\xi}-\vec{w}\|_\lambda^2 = \sum_i^M \lambda_i(\xi_i - w_i)^2$, where $\lambda_i \geq 0$ and $\sum_{i=1}^M \lambda_i = 1$. The effect is an axes-parallel scaling of the data according to the best discriminating feature dimensions, yielding to an ellipsoidal shape of a set of points equidistant to a prototype. This concept, further extended to an $M \times M$ matrix of relevance-factors (Generalized Matrix LVQ, GMLVQ [7]) yields $d(\vec{\xi},\vec{w}) = (\vec{\xi}-\vec{w}^p)^T\Lambda(\vec{\xi}-\vec{w}^p)$, where Λ is positive (semi-)definite and $\sum_{i=1}^M \Lambda_{i,i} = 1$. Hence the distance computation is shaped to a rotated ellipsoidal by accounting for correlations of the feature dimensions in the second diagonal elements of Λ. Both adaptive metrics enable non-linear decision boundaries by an extension to local relevance-vectors/matrices λ_p, Λ_p specific for every prototype, called localized GMLVQ/GRLVQ (LGMLVQ/LGRLVQ). Using stochastic gradient descent (introduced by GLVQ), the prototypes $\vec{w}^p$ of the network as well as the relevance-factors λ are updated by the derivatives $\partial E/\partial w$ and $\partial E/\partial \lambda$ of the classification error $E = \sum_{\vec{\xi}^{x,y}} f((d_J - d_K)/(d_J + d_K))$, $f(x) = (1 + exp(-x))^{-1}$. E must be minimized on the training data while maximizing the margin $d_K - d_J$, where d_J is the distance between $\vec{\xi}^{x,y}$ and the most similar prototype from the correct class $c(\vec{\xi}^{x,y}) = c(\vec{w}^J)$ and d_K is the distance to the most similar prototype from a wrong class. For figure-ground segregation a setup with two classes $C = 2$ is used where $c(\vec{w}^p) \in \{0..C-1\}$ encodes the class-membership of every weight/feature vector to figure or ground. Using similarity-based classification to compute the activation maps V_p as before, the final mask $\mathcal{A}$ is combined by choosing the maps from prototypes assigned to the foreground $\mathcal{A} = \sum_p^N c(\vec{w}_p)V_p$, $c(\cdot) \in \{0,1\}$.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

3 Evaluation of segmentation quality

Figure 2: Example for artificial distortion of the ground truth data. From left to right the original image, ground truth $\mathcal{G}$, distorted hypothesis $\mathcal{H}$ (patchsize $s_1 = 12$, shift $s_2 = 22$) and resulting segmentation $\mathcal{A}$.

Method	$G(\mathcal{A})$
GLVQ	0.09
GRLVQ	0.43
GMLVQ	0.50
LGRLVQ	0.60
LGMLVQ	0.92

Table 1: Average Similarity of foreground classification $\mathcal{A}$ to ground truth $\mathcal{G}$, where $\mathcal{H} = \mathcal{G}$.

The first evaluation of GLVQ using different metrics addresses the capability of foreground classification on the non-preprocessed feature-maps using the (noisy) $\mathcal{H}$ as supervised information. Therefore a database of rendered image sequences from 25 3D objects (bottles, boxes, cars etc.) is used. The arbitrarily rotated object-views are pasted in the centre of a typical scene (human in the background, hand near object), generated by tracking the view-centred hand in front of the camera systems. The available ground truth membership $\mathcal{G}$ of pixels to the foreground is used to generate artificial hypothesis maps $\mathcal{H}$ (Figure 2) by randomly selecting and shifting a patch from one position in the image to another (patch size s_1, shift distance s_2). The resulting foreground classification masks $\mathcal{A}$ are compared to the ground truth data using $G(\mathcal{A})$. In all experiments M=20 randomly initialized prototypes are used (6 figure, 14 ground) adapted by 10000 training-steps for each image with a learning-rate of 0.05/0.005 for the prototypes/relevance-factors.

We show in Table 1 the results for using the ground truth data $\mathcal{G}$ for supervised training, averaged over 700 images per object. Increasing complexity of the adaptive metrics from relevance-vectors to matrices and from global to local ones, which was the only modified parameter in this experiment, clearly leads to an increasing segmentation quality. Measured by the overlap $\mathcal{G}$, which considers only foreground-pixels, the resulting foreground mask reach an average similarity to the ground truth data up to 92% for LGMLVQ. Note that, although $G(\mathcal{A})$ can be very small, nevertheless the overall pixel classification performance is much better, e.g., 87% for GLVQ

$s_{1,2}$		$G(\mathcal{H})$	$G(\mathcal{A})$	$G(\mathcal{A}^*)$
0	0	1	0.92	0.95
8	8	0.90	0.84	0.87
10	16	0.78	0.76	0.79
12	22	0.69	0.73	0.77
12	30	0.62	0.73	0.77
14	38	0.54	0.74	0.78

Table 2: Average similarity to ground truth $\mathcal{G}$ of initial hypothesis $\mathcal{H}$ and results of LGMLVQ without ($G(\mathcal{A})$) and with post-processing ($G(\mathcal{A}^*)$).

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

and 98% for LGMLVQ, taking the background pixels into account. On the basis of results of Table 1, we further introduce additional noise to the training signal by using multiple degrees of distortions of $\mathcal{G}$ for supervised training of LGMLVQ (Tab. 2). This noise, as well as similar colors in foreground and background are responsible for overlapping clusters in feature-space. This problem cannot be solved, but the non-linear decision boundaries introduced by local transformations, as well as the even higher flexibility by using multiple prototypes for each class, allow a better representation of the heterogeneously structured data. Because of classification errors introduced by the method itself some reasonable distortion is required to observe the beneficial effect for our scenario ($G(\mathcal{A}) > G(\mathcal{H})$). But in this case, the higher model complexity enables a higher robustness to this distortion. Applying a closing operation ($\mathcal{A}^*$) to the more or less noisy $\mathcal{A}$, further improves results for succeeding processing.

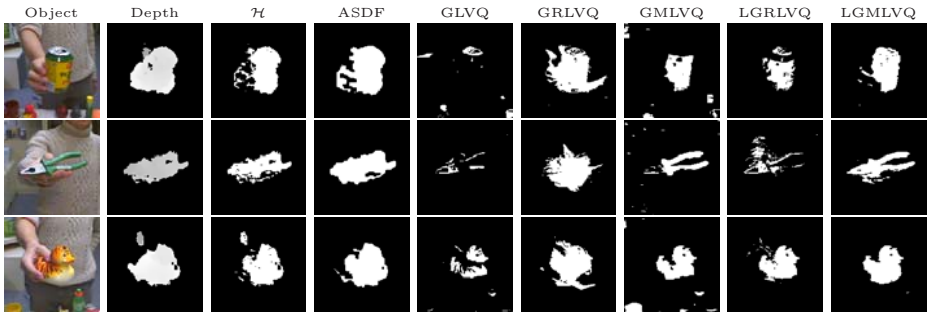
4 Evaluation of recognition performance

Finally we have to evaluate the capabilities of this approach on real image data and to investigate the effort of the derived object segmentations in the context of online object learning and recognition. Here we are using the data from [8] consisting of 50 natural, view centred objects with 300 training and 100 testing images without ground truth information. From the available depth and skin information the hypothesis $\mathcal{H}$ is computed, without additional prior information on object position (as used in [4], see Section 2). To compare the results of the different methods, the image regions defined by the foreground classification (i.e. the presented objects) are fed into a hierarchical feature processing stage [8]. For object learning and recognition a separate nearest neighbour classifier is applied to the derived high dimensional shape features. Figure 3 shows samples for $\mathcal{A}$ and the recognition performance from using the depth-map itself, hypothesis $\mathcal{H}$, the ASDF (used from [8]), and the results of the compared GLVQ-extensions. Despite of the the noisy data LGMLVQ is capable to represent (generalize) figure and ground on the basis of the best discriminating features, which enables a correct classification of the main object parts. Therefore using foreground classifications of LGMLVQ causes a significant improvement in recognition performance on real world data, whereas still running at reasonable time for online learning of 4 Frames/sec on a 3.6 GHz Intel Xeon processor machine.

5 Conclusion

In this paper we have compared several metrics extensions applied to GLVQ and finally adopt LGMLVQ in the domain of figure-ground segregation. In comparison to other metrics, we have shown that the extension to local matrices of relevance vectors leads to improved foreground classification resulting in a significant enhancement of object learning and recognition. Compared to the ASDF approach, which also directly addresses the foreground segmentation from an initial hypothesis, the supervised learning does not rely on additional a priori

assumptions about object position, size and segment-selection. From the viewpoint of supervised learning where the goal is a correct (pixel) classification, the incomplete approximation of $\mathcal{H}$ by $\mathcal{A}$ seems not desired. Here we use the advantage of GLVQ as large margin classifier, which accepts a small number of false classifications for more confident/robust decision boundaries, enabling a higher generalization to the underlying structures (e.g. object parts) in the image data.



Rec. Perf.:	73.5%	75.5%	77.8%	23.6%	31.6%	64.2%	68.7%	88.8%

Figure 3: From left to right: input image, depth-map, hypothesis $\mathcal{H}$ and derived $\mathcal{A}$ using GLVQ with Euclidian and adaptive metrics. Bottom row, the recognition performance of a separate nearest neighbour classifier on the segmented object images (300 images for training, 100 for testing). Observable is a gradual increase of segmentation quality and performance from taking into account correlations of the features as well as the usage of local transformation rather than global ones.

References

[1] H. Kim, E. Murphy-Chutorian, and J. Triesch. Semi-autonomous learning of objects. In *Proc. of the Conf. on Comput. Vision and Pattern Recogn. Workshop*, page 145, 2006.

[2] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1-2):41–63, 2005.

[3] S. X. Yu and J. Shi. Object-specific figure-ground segregation. In *Proc. of IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recogn.*, volume 2, pages 39–45, 2003.

[4] J. J. Steil, M. Götting, H. Wersing, E. Körner, and H. Ritter. Adaptive scene-dependent filters for segmentation and online learning of visual objects. *Neurocomputing*, 70(7-9):1235–1246, March 2007.

[5] A. Sato and K. Yamada. Generalized learning vector quantization. In *NIPS*, pages 423–429, 1995.

[6] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Netw.*, 15(8-9):1059–1068, 2002.

[7] P. Schneider, M. Biehl, and B. Hammer. Relevance matrices in LVQ. In *Similarity-based Clustering and its Application to Medicine and Biology*, 2007.

[8] H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. J. Steil, H. Ritter, and E. Körner. Online learning of objects in a biologically motivated visual architecture. *Int. Journal of Neural Systems*, 17(4):219–230, 2007.